未来予測シナリオ調査におけるテキストクラスタリングに関する研究

野田, 裕二 / NODA, Yuji

(出版者 / Publisher)
法政大学大学院理工学・工学研究科
(雑誌名 / Journal or Publication Title)
法政大学大学院紀要. 理工学・工学研究科編 / 法政大学大学院紀要. 理工学・工学研究科編
(巻 / Volume)
55
(開始ページ / Start Page)
1
(終了ページ / End Page)
5
(発行年 / Year)
2014-03-24
(URL)
https://doi.org/10.15002/00010493

未来予測シナリオ調査における テキストクラスタリングに関する研究

A STUDY ON TEXT CLUSTERING METHOD IN FUTULE PREDICTION SCENARIO STUDY

野田 裕二
Yuji NODA
指導教員　藤井 章博

法政大学大学院工学研究科情報電子工学専攻修士課程

In this paper, I propose In future prediction scenario study. In this study, it is the purpose of creating own indicators to measure the degree of similarity of the document. By performing an automatic document clustering specializing in future prediction survey conducted by National Institute of Science and Technology Policy, it is intended to realize the extraction of documents, relevant.
Key Words *:* clustering*,* information filtering *, TF-IDF,NISTEP*

1.　はじめに

近年,インターネットの急速な発達,そしてスマートフォンの爆発的な普及もあり,Web 上に存在する情報の価値はますます高まってきている.また,計算機システムの高性能化とネットワーク化にともない,膨大な電子化された情報が計算機上でアクセス可能になってきている.さらに,Web 上にどんどん蓄積されていく最新の情報だけでなく,過去に蓄積された情報の解析というのも,近年の企業活動においてますます重要な価値を生み出している.このような状況である近年,情報検索は,これらの膨大な情報の中から必要な情報を見つけ出すために必要不可欠な技術である.しかし,ここにおける情報の解析手法というのは様々なものがあり,解析対象とする情報の分野によっては,解析結果に偏りが現れる場合がある.本研究の目的は,この対象とする分野をあらかじめ絞り,その中における情報の持つ意味を解析し,対象とした情報がどの分野に属する情報なのか指標を作成し,導き出すものである.

2.　究背背景

（１）情報フィルタリング

インターネットの発達により,膨大な情報にアクセスが可能となったが,現在の情報の供給量は我々の必要としている情報量をはるかに上回っており,情報爆発（information explosion）あるいは情報洪水（information inundation）とも呼ばれる状況を作り出している.

情報検索は,我々の情報要求に応える技術であり,膨大な情報の中から必要な情報を能動的に取り出すためにはきわめて有用である.しかし一方で,情報要求がないにもかかわらず,否応なしに次々と我々のもとへ流れてくる情報も数多く存在する.このような情報の中から必要な情報のみを取り出し,不必要な情報を削除するための技術が情報フィルタリング（information filtering）である.

（２）文書の自動分類

情報アクセスを支援するひとつの手段として,たとえば文章をその内容に応じて分類・整理するということが考えられる.大規模な文書集合を分類・整理するためには,計算機による自動的な処理が必要となる.このための技術が文書の自動分類（document categorization, document classification）である.なお,文書の自動分類は大きくふたつに分けることができる.ひとつは,与えられた文書の内容があらかじめ設定されているカテゴリ（例えば,政治,経済,科学など）のいずれに属するかを決定するものであり,もうひとつは,類似した文書をグループ化（クラスタリング）することにより文書集合全体をいくつかのグループに分割するものである.

文書の自動分類は,ふたつの文書間の類似度,あるいは文書とカテゴリとの間の類似度を計算することに帰着できるので,基本的には情報検索と類似度の技術により実現することができる.ベクトル空間モデルに基づく方法,確率モデルに基づく方法,規則に基づく方法,識別学習に基づく方法などを始めとして,非常に多くの文書分類の方法が提案されている.

（３）情報抽出

情報検索が膨大な情報の中からユーザの要求に合った情報を能動的に取り出すのに対し,あらかじめ指定した情

報のみを効率的に取り出す技術が情報抽出（information extraction）である.すなわち,情報検索により取得した情報の中から,特に決まった情報のみを効率的に参照したい場合に情報抽出の技術が役に立つ.

日本語文章を対象にした情報抽出技術としては,固有表現の前後に出現する単語情報（表記,品詞）をコーパス（言語データベース）から学習して固有表現を抽出する研究,表層的な表現を手がかりとして電子メイルや電子ニュース記事のスケジュールやイベント情報を抽出する研究,Web ページや電子メイルを対象にして新製品情報を抽出する研究などがある.特に,抽出した固有表現をテンプレートに埋め込む際に,固有表現の同一性や多義性を認識することが,今後の技術的な課題であると言われている.

（４）本研究の概要

本研究の目的は、科学技術・学術政策研究所（NISTEP）が未来予測調査を行なうにあたり実施した専門家へのアンケート調査を基にした４７の分野への情報（未知のドキュメント Dα）の自動クラスタリングである.以下の図 1 に本研究の簡単な流れを示す.

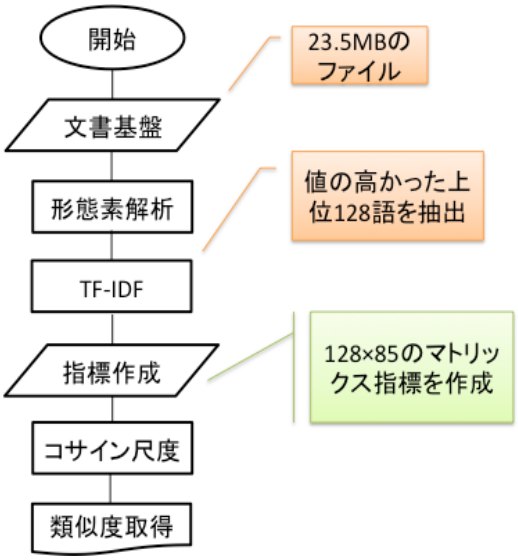

図１　作業の流れ

文書基盤として,科学技術・学術政策研究所（NISTEP）が未来予測調査を行なうにあたり実施した専門家へのアンケート調査のデータ 23.5MB を用い,各文書においてテキスト解析を行い,TF-IDF 値が最も高かった上位 128 語を抽出する.その 128 語を基にして,128×85 のオリジナルのマトリックス指標を作成する.この指標を使用し,ターゲット文書との類似度をコサイン尺度を用いて計算し,関連度を導き出す.なお,今回の研究ではターゲット文書として,はてなブックマークから『宇宙科学』に関するブログをクローラーで抽出し,解析対象としている.

a) 対象としたデータ

本研究では,科学技術・学術政策研究所が行なったシナリオ調査における全 47 分野,85 文書(23.5MB)を用いて,ターゲットとした文書がこの 47 分野のうちどの分野に最も近しいかを検証している.科学技術・学術政策研究所は,国の科学技術政策立案プロセスの一翼を担うために設置された国家行政組織法に基づく文部科学省直轄の国立試験研究機関であり,行政ニーズを的確にとらえ,意思決定過程への参画を含めた行政部局との連携,協力を行うことが期待されている機関である.

b) 指標の作成の流れ

以下の図２に指標作成の流れを示す.

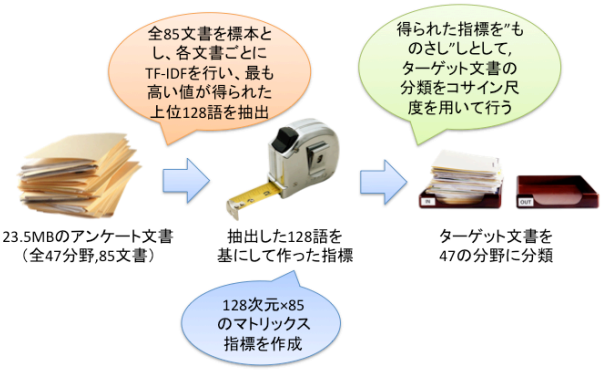

図２　指標（ものさし）作成に関する流れ

3.　索引語の抽出と重み付け

情報検索においては,しばしば文書の内容を文書中に含まれる単語の集合で近似するということが行なわれる.しかし,文書に含まれている単語全てが一律に文書の内容と関係しているわけではない.

（１）索引語の抽出

日本語には単語間に明示的な区切りを入れるという習慣がないため,単語を抽出するためには形態素解析を行う必要がある.しかし,日本語では単語とは何かという定義が明確でない部分があり,同一の文字列に対しても多様な単語分割の可能性がある.銅のように単語分割が行なわれるかは,使用した形態素解析のシステムに依存するし,場合によっては,同じシステムで同一の文字列を形態素解析する際にも,文脈（前後にある単語）により異なった分割が行なわれる可能性もある.

（２）索引語の重み付け

n個の文書$D_1, D_2, \cdots, D_n$があり,これらの文書集合から全部でm個の索引語$w_1, w_2, \cdots, w_m$が抽出されたとする.このとき,索引語w_iの文書D_jにおける重みd_{ij}は,一般に以下のような 3 つの指標によって特徴付けることができる.

a) 局所的重み（local weight）：l_{ij}

索引語w_iの文書D_jにおける出現頻度に基づき計算される重みである.主に再現率向上を目的とした重みであり,文書中に頻繁に出現する索引語に対して大きな値が与えられる.

b) 大域的重み（global weight）：g_i

文書集合全体にわたる索引語w_iの分布を考慮して,決定される重みである.主に適合率向上を目的とした重みであり,特定の文書に集中して出現する索引語に対して大きな

値が与えられる.

c) 文書正規化係数（document normalization factor）：n_j

文書が長くなるにつれ,その中に含まれる索引語の数も増えるため,長い文書に含まれる索引語のほうが大きな重みを持つようになる.文書正規化係数は,文書の長さによる影響をなくす目的で導入するものである.

上記の3つの指標から,以下の式により索引語の重みd_{ij}を決定する.

$$d_{ij} = \frac{l_{ij} g_i}{n_j} \qquad (4.1)$$

（３）特徴語の抽出

文書間の特徴語の抽出方法として TF-IDF という手法がある.

a) TF（Term Frequency）

TF とは, ドキュメント内でキーワードがどれだけ多く仕様されているのかを示す指標として用いられる. キーワードを多く含むドキュメントほど, そのキーワードについて詳しく説明しているものと考える. TF は以下の式で表される.

$$TF = \frac{ドキュメント内のキーワード出現回数}{ドキュメント内の総単語数}$$

b) DF（ Document Frequency）

DF とは, そのキーワードがどれだけの数のドキュメントで使用されているのかを示す指標である. 多くのドキュメントで使用されているキーワードより, 少ないドキュメントで使用されているキーワードの方がそのドキュメントの特徴をよく表すものと考える. DF は以下の式で表される.

$$DF = \frac{総ドキュメント数}{キーワードが出現するドキュメント数}$$

c) TF-IDF(Term Frequency – Inverse Document Frequency）

TF-IDF とは主に情報検索で最も有名で広く利用されている重み付けの手法の一つである.IDF とは DF の逆数を表すので,

$$(TF * IDF) = \frac{ドキュメント内のキーワード出現回数}{ドキュメント内の総単語数} log \frac{総ドキュメント数}{キーワードが出現するドキュメント数}$$

と表せる.

$$TF・IDF =$$

5　ベクトル空間モデルに基づく文書検索

ベクトル空間モデル（vector space model）は,1970 年代にサルタン（Salton）らによる SMART システムで採用されたモデルであり,情報検索における代表的な検索モデルとして知られている.ベクトル空間モデルの最大の特徴は,文書および検索質問を多次元のベクトルという中間的な媒体により表現し,ベクトル間の類似度を計算することにより文書の類似度を実現している点にある.

（１）ベクトルの演算

ベクトル$x = [x_1\ x_2\ \cdots\ x_n]^T$と$y = [y_1\ y_2\ \cdots\ y_n]^T$に対し,$x$のノルム（norm）,および$x$と$y$の内積（inner product）を次のように定義する.ノルムは,ベクトルを有効直線で表した時の長さを表している.

a)　ノルム

$$||x|| = \sqrt{\sum_{i=1}^{n} x_i^2} = \sqrt{x_1^2 + x_2^2 + \cdots + x_n^2} \qquad (5.11)$$

b)　内積

$$\mathrm{x} \cdot \mathrm{y} = \sum_{i=1}^{n} x_i y_i = x_1 y_1 + x_2 y_2 + \cdots + x_n y_n \qquad (5.12)$$

（２）ベクトル空間モデル

検索対象となる文書$D_1, D_2, \cdots, D_n$とし,これらの文書集合全体を通して全部でm個の索引語$w_1, w_2, \cdots, w_m$があるとする.このとき,文書D_jは,次のようなベクトルで表現される.これを文書ベクトル（document vector）と呼ぶ.

$$d_j = \begin{bmatrix} \mathrm{d_{1j}} \\ \mathrm{d_{2j}} \\ \vdots \\ \mathrm{d_{mj}} \end{bmatrix} \qquad (5.21)$$

ここで,d_{ij}は検索語w_iの文書D_jにおける重みである.

また,文書集合全体は,次のようなm×n行列Dによって表現することができる.

$$D = [d_1\ d_2\ \cdots\ d_n] = \begin{bmatrix} \mathrm{d_{11}}\ \mathrm{d_{12}} \cdots \mathrm{d_{1n}} \\ \mathrm{d_{21}}\ \mathrm{d_{22}} \cdots \mathrm{d_{2n}} \\ \vdots\ \ \vdots\ \ \ddots\ \ \vdots \\ \mathrm{d_{m1}} \mathrm{d_{m2}} \cdots \mathrm{d_{mn}} \end{bmatrix} \qquad (5.22)$$

行列Dを索引語・文書行列（term-document matrix）と呼ぶ.索引語・文書行列の各列は文書に関する情報を表しているベクトルであるが,同様に,索引語・文書行列の各行は索引語に関する情報を表しているベクトルであり,これを索引語ベクトル（term vector）と呼ぶ.

検索質問も,文書と同様に,索引語の重みを要素とするベクトルで表現することができる.検索質問文に含まれる索引語　w_iの重みをq_iとすると,検索質問ベクトルqは次のように表される.

$$q = \begin{bmatrix} q_1 \\ q_2 \\ \vdots \\ q_m \end{bmatrix} \qquad (5.23)$$

実際の文書検索においては,与えられた検索質問文と類似した文書を見つけ出す必要があるが,ベクトル空間モデルでは,これを検索質問ベクトルqと各文書ベクトルd_jの間の類似度を計算することにより行なう.ベクトル間の類似度の定義としてはさまざまなものが考えられるが,文書検索においてよく用いられているものはコサイン尺度（2つのベクトルのなす角度）や内積である.

a) コサイン尺度

$$\cos(d_j, q) = \frac{d_j \cdot q}{\| d_j \| \| q \|} = \frac{\sum_{i=1}^{m} d_{ij} q_i}{\sqrt{\sum_{i=1}^{m} d_{ij}^2} \sqrt{\sum_{i=1}^{m} q_i^2}} \quad (5.24)$$

b) 内積

$$d_j \cdot q = \sum_{i=1}^{m} d_{ij} q_i \quad (5.25)$$

なお,ベクトルの長さが 1 に正規化（コサイン正規化）されている場合には,コサイン尺度と内積は一致する.

6　評価

（１）TF-IDF 値の高かった上位 128 語を抽出

以下の表 1 に,科学技術・学術政策研究所が行なった未来予測シナリオ調査で実施した 47 の分野の専門家へのアンケート調査を基にした全85 文書を解析し,TF-IDF 値が最も高かった単語上位 128 単語を抽出した結果を示す.

表 1 TF-IDF 値の上位 128 語抽出結果

i	Words	Values
1	エンタテインメント	0.159
2	量子	0.149
3	感染症	0.134
4	数学	0.114
5	宇宙	0.104
6	金融	0.094
7	情動	0.092
8	cad	0.087
⋮	⋮	⋮
128	半導体	0.031

（２）抽出した 128 語を用いて指標を作成

以下の図 3 に, 抽出した 128 語を用いて作成した 128 次元×85 の多次元ベクトル指標を示す.

(128 Words \ Documents)	d_1	d_2	d_3	d_4	d_5	…	d_{85}
エンタテインメント	0	0	0	0	0	…	1
量子	0	0	0	0	0	…	1
感染症	0	0	0	0	0	…	0
数学	0	0	0	1	1	…	0
宇宙	0	0	0	0	0	…	0
⋮	⋮	⋮	⋮	⋮	⋮	⋱	…
半導体	0	0	0	0	0	…	0

図 3 抽出した 128 語を用いて作成した指標

（４）作成した指標を使ったクラスタリング結果

作成した指標を用いて,未知の文書 D α がどの分野へクラスタリングされているかを検査した.以下の表にその結果を示す.ここで未知の文書 D α は,結果をわかりやすくする目的のため,クローリングによって取得した『宇宙科学』分野の文書とした.

表 2 未知の文書 D α における類似度評価

類似順位	文書の分野	類似度
1	宇宙科学D1	0.666
2	宇宙科学D2	0.586
3	衛星技術D1	0.402
4	衛星技術D2	0.378
5	地球深部探査	0.316

（５）宇宙科学に関するブログの類似度結果

今度は作成した指標を用いて,クローリングによって取得したはてなブックマークにおける“宇宙科学”分野のブログ文書 10 個（T1～T10）に対し,コサイン尺度を用いて類似度の計算を行なった.以下の表 3 にその結果を示す.

表 3 宇宙科学に関するブログの類似度

類似順位	ブログ	類似度
1	T4	0.665
2	T7	0.624
3	T2	0.621
4	T9	0.602
5	T3	0.584
6	T1	0.579
7	T5	0.502
8	T6	0.403
9	T8	0.389
10	T10	0.332

7　考察

ある未知のブログの文書 Dα を解析した場合、類似度評価の結果から宇宙科学とそれ以外の分野である一定の差が得られた.このため,科学技術分野の文書に対する未知の文書 D α の文書の分類が正しくできていると評価した.

この結果から,文書分類という技術的な課題は達成した.しかし,次の宇宙科学に関する内容かを調べた結果の類似度からは,類似度が 0.4 に満たない数値でも調査対象とした４７分野の中では一番高い数値が得られた.ブログの内容は調査対象である47の中では宇宙科学の内容が高かったということも考えられる.よって,今後は類似度の値が低くても内容が他分野と共有していないかといった確認が必要である.

8　むすび

本研究では,分類対象とする分野を科学技術・学術政策研究所が行なったシナリオ調査においての専門家からのアンケート調査をもとにした全47分野への科学技術分野への分類を行なった.オリジナルの指標（ものさし）を作成し,それを用いて一定の評価が得られたことは,作成した分類機が性格に動作していることが確認できた.しかし,結果から類似度の値が低かったが他の分野の類似度が更に低い値だったために,結果的には文書の分野が宇宙科学として選ばれているものもあった.今後はそういった類似度の低かったものに対して,更に別の方法で文書の類似度を確認する必要性があると考える.

謝辞：本研究を進めるにあたりご指導を頂いた藤井章博准教授, ならびに知恵をご教授頂いた多くの先生方, そして研究活動を共にした研究室の仲間や産学共同研究で関わった皆様に心から感謝を申し上げます.

参考文献

[1] Satnam Alag. Collective Intelligence in Action. Manning Publications Co. 2008.

[2] Jurgen Umbrich and Andreas Harth and Aidan Hogan and Stefan Decker Four Heuristics to Guide Structured Content Crawling.

[3] 北研二. 津田和彦. 獅々堀正幹. 情報検索アルゴリズム. 共立出版株式会社.2005.

[4] Katsuji Bessho. Osamu Furuse. Ryoji Kataoka. Concept Vector Generation Method Based on Co-occurrences between Words and Semantic Attributes . NTT Cyber Solutions Laboratories.NTT Corporation.2006

[5] Taiichi Hashimoto. Koji Murakami. Takashi Inui. Kazuo Utsumi. Masamichi Ishikawa. Document Clustering for Social Problem Detection and Cluster Evaluation Measures